Similar problem statement at Topcoder.Com

**Objective:** To develop language classifier that correctly identify the languages spoken in audio recordings. Given speech data from multiple languages, the classifier must identify which languages are spoken. The competition asks for three best possible spoken languages per test audio files.

**URL for problem statement:**

<https://community.topcoder.com/longcontest/?module=ViewProblemStatement&rd=16555&pm=13978>

**Dataset:** It has 66,176 speech files (.mp3) as training dataset of 176 languages, each language has 376 samples of 10 sec. And has 12,320 speech files (.mp3) as test dataset of 10 sec.

Unfortunately, it **doesn’t have Nepali language** speech files.

Please find below the languages available in this dataset.

| **Language** | **Number of Samples** |
| --- | --- |
| Adiokrou | 376 |
| Aguaruna Awajun | 376 |
| Akha | 376 |
| Akoose | 376 |
| Alangan | 376 |
| Albanian Tosk | 376 |
| Alladian | 376 |
| Altai Southern | 376 |
| Anufo | 376 |
| Arabic | 376 |
| Arabic Egyptian Spoken | 376 |
| Arabic Sudanese Spoken | 376 |
| Aruamu | 376 |
| Ashaninca | 376 |
| Avokaya | 376 |
| Awa PNG | 376 |
| Awadhi | 376 |
| Aymara | 376 |
| Azerbaijani North | 376 |
| Bali | 376 |
| Bekwarra | 376 |
| Belize Kriol English | 376 |
| Bench | 376 |
| Bhojpuri | 376 |
| Biate | 376 |
| Bicolano Central | 376 |
| Bokyi | 376 |
| Bora | 376 |
| Bru Eastern | 376 |
| Cakchiquel South Central | 376 |
| CandoshiShapra | 376 |
| Carib | 376 |
| Chamacoco | 376 |
| Chayahuita Shawi | 376 |
| Chinantec de Usila | 376 |
| Chipaya | 376 |
| Chuj Ixtatan | 376 |
| Cofan | 376 |
| Dan East | 376 |
| Dangaleat | 376 |
| Dinka Northeastern | 376 |
| Duala | 376 |
| Duri | 376 |
| Dutch | 376 |
| Eleme | 376 |
| Ese Ejja | 376 |
| Fai South | 376 |
| Galela | 376 |
| Gamo | 376 |
| Garhwali | 376 |
| Gidar | 376 |
| Guahibo | 376 |
| Guayabero | 376 |
| Hadiyya | 376 |
| Hiligaynon | 376 |
| Hindi | 376 |
| Hmong Daw | 376 |
| Huambisa | 376 |
| Huave de San Mateo del Mar | 376 |
| Huli | 376 |
| Ignaciano | 376 |
| Ilokano | 376 |
| Ipili | 376 |
| Ixil Nebaj | 376 |
| Javanese | 376 |
| Jingpho | 376 |
| Jola Kasa | 376 |
| Jur Modo | 376 |
| Kalabari | 376 |
| KalmykOirat | 376 |
| Kannada | 376 |
| Karakalpak | 376 |
| Karamojong | 376 |
| Kashinawa | 376 |
| Kazakh | 376 |
| Kera | 376 |
| Khasi | 376 |
| Khmer Northern | 376 |
| Kiche Cunen | 376 |
| Kilivila Kiriwina | 376 |
| Kok Borok | 376 |
| Kolami Northwestern | 376 |
| Konso | 376 |
| Koorete | 376 |
| Korean North | 376 |
| Korean South | 376 |
| Kui | 376 |
| Kuman | 376 |
| Kumyk | 376 |
| Kyrgyz | 376 |
| Lampung Api | 376 |
| Lao | 376 |
| Lobi | 376 |
| Luwo | 376 |
| Madi | 376 |
| Maka | 376 |
| Malagasy Plateau | 376 |
| Malay | 376 |
| Marba\_Azumeina | 376 |
| Maya Mopan | 376 |
| Mazahua Central | 376 |
| Mentawai | 376 |
| Mixtec Ayutla | 376 |
| MofuGudur | 376 |
| Mongolian Halh | 376 |
| Mongondow | 376 |
| Morisyen | 376 |
| Moro | 376 |
| Mundari | 376 |
| Munukutuba | 376 |
| Murle | 376 |
| Musgu | 376 |
| Naga Ao | 376 |
| Nahuatl Guerrero | 376 |
| Nahuatl Highland Puebla | 376 |
| Nahuatl Northern Puebla | 376 |
| Napu | 376 |
| Nigerian Pidgin | 376 |
| Ojibwa Northwestern | 376 |
| Oromo Eastern | 376 |
| Otomi Mezquital | 376 |
| Paez | 376 |
| Palauan | 376 |
| Palaung Ruching | 376 |
| Pamona\_Bahasa Taa | 376 |
| Pampangan | 376 |
| Parecis | 376 |
| Patamona | 376 |
| Paumari | 376 |
| Peve | 376 |
| Pijin | 376 |
| Pokoot | 376 |
| Polish | 376 |
| Qeqchi | 376 |
| Quechua Eastern | 376 |
| Quechua MargosYarowilcaLauricocha | 376 |
| Quechua North Junin | 376 |
| Quechua Southern Conchucos Ancash | 376 |
| Quechua Southern Pastaza | 376 |
| Romani Sinti | 376 |
| Romanian | 376 |
| Samoan | 376 |
| Sango | 376 |
| Sanuma | 376 |
| Sasak | 376 |
| Sauria Paharia | 376 |
| Sebat Bet Gurage\_Gurage Chaha | 376 |
| Shilluk | 376 |
| ShipiboConibo | 376 |
| Shuar | 376 |
| Tamasheq | 376 |
| Tektiteko | 376 |
| Tepehua Pisaflores | 376 |
| Tepehua Tlachichilco | 376 |
| Terena | 376 |
| Teribe | 376 |
| Teso | 376 |
| Thai | 376 |
| Themne | 376 |
| Ticuna | 376 |
| Tobelo | 376 |
| Tol | 376 |
| Totonac Highland | 376 |
| Trio | 376 |
| Tucano | 376 |
| Tuva | 376 |
| Tzeltal Bachajon | 376 |
| Vai | 376 |
| Vietnamese | 376 |
| Waimaha | 376 |
| Waiwai | 376 |
| Wichi Lhamtes Nocten Weenhayek | 376 |
| Yali Angguruk | 376 |
| Yawa | 376 |
| Zapotec Isthmus | 376 |
| Zapotec Sierra de Juarez | 376 |
| **Grand Total** | **66176** |

**Solution:**

Topcoder didn’t provide the winning solution, however the participant shared their GitHub link and written blogs about it. I found two winners solution(rank 1 and rank 10).

**Rank 1**

He has used Gaussian Mixture Model to trained the data and used logistic regression as final step to calibrate the individual GMM's prediction

<https://github.com/CatalinTiseanu/spoken-language-identification>

<https://github.com/CatalinTiseanu/spoken-language-identification/blob/master/SpokenLanguages2%20-%20report.pdf>

Main points about his solution:

1. Use sox to convert the mp3’s to wav
2. Use sphinx\_fe to extract 13 Mel Frequency Cepstral Coefficients (MFCC) features
3. Train a Gaussian Mixture Model with 2048 mixtures for each language feature matrix.
4. Finally, calibrate the individual language predictions using Logistic Regression

**Rank 10**

Rank 10 winner who had acclaimed 95% accuracy in training data. Please find the solution and code at Github.

<https://yerevann.github.io/2015/10/11/spoken-language-identification-with-deep-convolutional-networks/>

Github Link.

<https://github.com/YerevaNN/Spoken-language-identification>

The main points about his solution is.

1. Used Deep Convolutional Network (Machine Vision).
2. Converted mp3 to wav and created spectrogram of a wav file.
3. Spent lot of effort on learning rate and data augmentation.
4. Used ensemble as final submission.

He had not used RNN/LSTM model or any hybrid models (neural network and HMM). However, these could improve the accuracy.